AMERICAN JOURNAL OF PHYSICAL ANTHROPOLOGY 100:207–224 (1996)

"Race" Specificity and the Femur/Stature Ratio

MARC R. FELDESMAN AND ROBERT L. FOUNTAIN
Department of Anthropology (M.R.F.) and Department of Mathematical Sciences (R.L.F.), Portland State University, Portland, Oregon 97207

KEY WORDS Stature estimation, Forensics, Statistical validation, Body size, Fossil hominids

ABSTRACT This inquiry explores a series of problems related to the femur/stature ratio first raised by Feldesman et al. (1990). In particular, we used a revised data set and a more elaborate research protocol to address questions pertaining to: (1) whether the femur/stature ratios of three quasi-geographic "races" ("Blacks," "Whites," "Asians") are statistically significantly different; (2) whether these are statistically (as opposed to biologically) coherent groups; and (3) whether the "race"-specific ratios are more accurate than the simple generic femur/stature ratio.

We used ANOVA, ANOCOVA, post hoc analysis, *k*-means cluster analysis, linear discriminant functions, and approximate randomization to determine whether the group differences in the ratio were significant, and to assess the coherence of the "racial" groups themselves. We used validation procedures including mean absolute deviation, mean squared error, and Pitman's measure of closeness of a known sample of 798 femur/stature pairs to compare the accuracy of the generic ratio and the group-specific ratios.

The results confirmed that the "Black" femur/stature ratio is significantly different from those of "Whites" and "Asians"; however, group coherence was poor, with results barely better than chance.

We found that "race"-specific ratios slightly outperform the generic ratio when "race" is certain, but the gains are small for the assumptions required. More significantly, however, we found that when "race" attribution is uncertain or unknown, as in paleoanthropology, the wrong ratio (or the wrong regression equation) performs poorer than the generic femur/stature ratio. As a result, we recommend that researchers continue using the generic femur/stature ratio to estimate stature in pre- and protohistoric populations. An alternative equation, a generic regression, yields even better stature estimates; however, we urge further study before recommending that researchers use this instead of the more thoroughly tested generic femur/stature ratio.

Problems associated with estimating stature in living and fossil humans continue to interest physical anthropologists. In 1990, Feldesman, Kleckner, and Lundy (hereinafter FKL) argued that a simple ratio of femur length to stature yielded remarkably reliable estimates of stature in contemporary forensic specimens and in fossil hominids (i.e., mid- to late Pleistocene forms) for which neither gender nor race could be ascertained with certainty. Their discovery of the ratio was not new (Sué, 1775), nor was its application to estimating stature of fossil hominids (Reed and Falk, 1977) or contemporary populations original (e.g., Hrdlicka, 1972). FKL were, however, the first to systematically evaluate

Received April 20, 1995; accepted December 26, 1995.

Address reprint requests to Marc R. Feldesman, Department of Anthropology, Portland State University, P.O. Box 751, Portland, OR 97207-0751.

its broad applicability to both forensic and paleontological stature estimation problems. Whether by coincidence, by inspiration, or simply as a continuation of the general interest in the whole problem of stature estimation, a spate of new studies of stature estimation in living and fossil hominids followed (Feldesman, 1992a,b; Formicola, 1993; Holland, 1992, 1995; Jantz, 1992; McHenry, 1991; Peterson, 1992; Ruff and Walker, 1993; Sciulli et al., 1990; Sciulli and Giesen, 1993; Sjøvold, 1990). These papers address a multitude of issues pertaining to the problems of stature estimation in living and fossil hominids, and from them a series of new issues emerges. It is not possible to address all these papers at the present time; instead, we confine our review to those addressing conclusions reached in FKL.

Sjøvold (1990) developed gender- and "race"-neutral major axis regression equations ("the weighted line of organic correlation") to estimate stature from long bones of living and prehistoric humans. He reported that he obtained more accurate results with this technique over the broad range of hominid statures than he did with Model I regression-based stature-estimating equations, as well as the FKL ratio procedure.

McHenry (1991) used FKL's method to reconstruct stature from femur lengths of Plio-Pleistocene fossil hominids, both updating and revising his 1974 study on the same subject. He found considerable stature dimorphism in both *Australopithecus afarensis* and *Homo habilis*, smaller than expected stature in "robust" australopithecines, considerably less size dimorphism in other Plio-Pleistocene species, and weaker evidence for the common belief that hominids got progressively taller throughout their evolutionary history.

Feldesman (1992b) developed femur/stature ratios for "White" juvenile males and females between the ages of 12 and 18. He found that, between the ages of 12 and 18, the juvenile ratios are significantly different from the adult ratio, that age variations are not significant, and that gender differences are pronounced. Thus gender-specific ratios are required to estimate stature over this age range. He also discovered that juvenile ratios yielded better estimates of stature when applied to "White" juveniles than do the "race"- and gender-matched adult regression equations (Trotter and Gleser, 1958), which significantly overestimate juvenile stature. As a result, when the juvenile male ratio is applied to the *Homo erectus* fossil (WT 15000) from Nariokotome, it yields a significantly shorter stature than the Trotter and Gleser adult regression equations (contra, Leakey and Walker, 1985; and pro, Ruff and Walker, 1993).

Formicola (1993) compared a variety of different techniques, including the Trotter and Gleser formulae, as well as formulae from Pearson (1899), Bach (1965), Olivier et al. (1978), Sjøvold (1990), and FKL to estimate stature from limb bones of Neolithic hominids spanning seven European countries. He found that the obvious choices of stature-estimating formulae (e.g., Trotter and Gleser's "White" equations or Bach's "German" equations) gave poor stature estimates in these specimens. Instead, Formicola found that Pearson's (1899) equations and, surprisingly, Trotter and Gleser's (1952) "Black" formulae yielded the "best" stature estimates for these European samples. The FKL femur/stature ratio and Sjøvold's (1990) "line of organic correlation" gave credible estimates for shorter and mid-range statures, but tended to overestimate stature in tall individuals.

Ruff and Walker (1993) devised a thermoregulatory model and argued for a climate-based choice of stature-estimating equations in fossil hominids, criticizing FKL's technique in the process. They based stature estimates for the juvenile *Homo erectus* from Nariokotome, Kenya (WT 15000) on published forensic equations developed for contemporary African populations (Lundy and Feldesman, 1987; Feldesman and Lundy, 1988; Allbrook, 1961) who live in climates similar to those inhabited by these African hominids.[1]

[1]Ruff and Walker's (1994) thermoregulatory argument yields credible estimates of WT 15000's at-death stature; however, there is no epistemological basis for Ruff's (1993:58) claim that these are ". . . the most reasonable estimates." Estimates virtually identical to those offered by Ruff and Walker for WT 15000 were obtained by Feldesman (1992a) from femur/stature ratios derived for juvenile male White Europeans between the ages of 12 and 18. The congruence of results between these two studies suggests that the height estimates are probably fairly close to "true" stature, but whether this is due to coincidence, developmental factors, thermoregulation, or some other unknown factor cannot be determined without additional information (e.g., an adult male skeleton whose morphology is presaged by WT 15000, and/or another more complete skeleton like WT 15000).

These studies consistently show that the forensic equations for contemporary humans have limited or questionable accuracy when applied to prehistoric humans. Furthermore, some of these recent results have shown that when modern regression equations are applied to appropriately matched contemporary samples, the results suffer moderate to significant inaccuracies when compared to other techniques. Numerous different strategies have been developed to resolve these problems, with varying degrees of success. In the main, they include alternate equations and equation forms that are less gender and population specific.

In 1990, FKL used the alternate femur/stature ratio method to estimate stature. Even with a small sample, they found unequivocal statistical evidence that this ratio was not significantly different in males and females. That finding prompted them to consolidate all data and develop a single, gender-undifferentiated ratio. The "race" issue was not as clear and unambiguous. The initial sample of 51 was subdivided into three "races," whose boundaries were drawn partly along geographic lines, but whose membership corresponded more or less to generally accepted historical evidence of genetic propinquity. The small number of data points prevented subdividing the sample into more than three groups—"Black," "White," and "Asian." Excluded was any group whose genetic background was confounded by a well known history of admixture (e.g., US Mexicans, US Puerto Ricans, and "Russians"); however, FKL were well aware that most of the populations included count not be regarded as genetically homogeneous or geographically contiguous. Nevertheless, they performed a one-way analysis of variance based on 46 cases spread unevenly over the three "racial" groups. They found significant differences among the mean femur/stature ratios, and subsequent testing showed it to be the "Black" femur/stature ratio that was responsible for the significant F-test. Neither "Whites" nor "Asians" could be distinguished from one another. While FKL reported this result, they were ambivalent about its significance. The results were not unexpected in a broad climatic and geographic sense, but FKL felt that more data were required before they could make a definitive recommendation about the use of such "race"-specific ratios. In particular, FKL argued that such a recommendation would be mooted by a common fact of paleontological life: it is extremely difficult, if not impossible, to make decisions about "racial" affinity in hominid fossils. Therefore, it did not make a lot of sense to recommend formulae whose use would be limited by the very material upon which they were supposed to be applied. Consequently, little emphasis was placed on the finding of "race" specificity in the ratio.

Ruff (especially 1993, but also 1991 and 1994) criticized FKL's decision to minimize the "race" specificity of the ratios, and argued that while genetic or geographic "race" was not the issue, the connection between geography and the more important thermoregulatory and ecological factors could not be ignored in choosing stature-estimating equations for fossil hominids. Thus, he argued that while "race" was not a compelling factor, body proportions were influenced by climate, and climate is approximated by geography—a significant consideration in formulating geographic "races." Specifically, Ruff suggested (1993:58) that stature estimates for fossil hominids should be based on equations from populations that share ". . . thermoregulatory similarity resulting from . . . similar climatic environments." Ruff (1994:72) further noted that ". . . stature must be estimated with skeletal material . . . using appropriately proportioned living reference samples." FKL anticipated Ruff's first criticism and agreed with him in principle, but argued that the approach is not practical because there is a serious deficiency of estimating equations for populations sampling the world's major climatic zones. Ruff's second criticism, not formally advanced at the time the FKL paper was published, seems to us to beg the question: we cannot use appropriately proportioned living reference samples to estimate stature in fossil hominids without knowing the proportions of those hominids. To do so requires that at some general level we know stature, the very quantity we are interested in estimating.

That regression equations developed for use in contemporary populations continue to be used in fossil hominids is quite surprising

in view of the problems that have been reported since the late 1980s. These modern regression equations are bound to contemporary "races" or populations, as well as to gender, and were developed for entirely different purposes many years ago. Furthermore, to use these equations on individuals whose bone lengths fall at the edge of, or outside the range of, the original population distribution violates a fundamental assumption in regression analysis, regardless of whether the body proportions are the same or not. Thus, it does not come as much of a surprise that the original estimate of WT 15000's stature (Leakey and Walker, 1985) was significantly higher than later estimates (Feldesman and Lundy, 1988; Feldesman et al., 1990; Feldesman, 1992a, Ruff and Walker, 1993), which are virtually identical despite vastly different procedures and philosophies for estimating stature. The original estimates were derived from regression equations for adult male American Blacks whose mean femoral lengths were more than two standard deviations larger than WT 15000's. The later estimates used reference populations whose mean femoral lengths were consistent with WT15000's femur length, or came from climatic/geographic zones more closely approximating the environment in which WT 15000 was reputed to have lived, or used modern white juvenile males of approximately the same age and stage of development.

Feldesman (1992a) recognized the validity of Ruff's thermoregulatory argument, but felt it was futile to try to completely disentangle all these sources of variation with limited data. Instead, he sought resolution to the question of whether FKL's 1990 aggregated "race"-specific results were both statistically meaningful and helpful in disentangling some of these issues. Feldesman saw this as an indirect way to address some of the issues that Ruff raised, since the initial groupings were partially based on geography and climatic factors. No additional data had become available to address this question. Therefore, Feldesman "manufactured" data to test the hypothesis that these broadly constructed "races" showed no proportional differences between the femur length and stature, despite expectations based on thermoregulatory principles and ecogeographic "rules." He used a computer simulation technique called the "bootstrap" (explained most cogently in Efron and Tibshirani, 1993) to resample the original data set and generate estimates of the true probability of "racial" differences under nonparametric conditions. Feldesman found that the 1990 results could be replicated in more than 95% of the bootstrapped samples; however, multiple comparison tests did not reveal the same or even a consistent pattern to the results. These results failed to provide a clear answer to the question of when and whether a "race"-specific ratio should be used in lieu of the "generic" ratio. Moreover, simple validation studies comparing the predicted statures of known individuals using "race"-specific ratios, generic ratios, and "race"- and gender-specific regression equations failed to confirm any one technique as "best" in any given situation.

Clearly, while estimating stature in living and fossil hominids continues to interest researchers, there is no clear resolution of the various problems addressed by any of the recent analyses, and no consensus on how to proceed. None of the recent students have successfully navigated the issue of the "race" specificity of stature or of stature-estimating equations. Furthermore, the matter of the comparative efficacy of the generic femur/stature ratio in relation to other estimators has not been settled to our or to others' satisfaction. While the influences of thermoregulation, climate, and body proportions on stature estimates are interesting and important matters, the results from the studies that have considered them are sufficiently contradictory that we have chosen not to address them directly here. Biology, paleoclimatology, genetics, and simple logic suggest they should be significant, but we are not so persuaded by the results of the various analyses that in the end they make much difference in estimating fossil hominid stature. This is not to say that they are not important and should be ignored, but instead present the researcher, faced with fragmentary and very incomplete hominid remains from times long past, a nearly impossible set of variables to resolve.

In evaluating the applications and criticisms of the original FKL paper, it became clear that the "race" specificity issue was a lingering problem. Before the femur/stature

ratio could be applied confidently to fossil populations, we needed to satisfy ourselves that the ratio gave sufficiently accurate answers in modern circumstances to be useful. We decided to focus our inquiry on issues pertinent to its use in modern populations, since no direct tests of its accuracy could ever be performed with incomplete fossil data. Thus we sought additional data and revised the analytical framework originally formulated in FKL to answer the following questions: (1) Do different "races," quasigeographically defined, show different femur length to stature proportions? (2) Do the "racial" groups cohere enough statistically on the basis of the femur/stature ratio to be useful classificatory categories? (3) If "race"-specific femur/stature ratios are valid, do they yield more accurate stature estimates than the "generic" ratio in cases where both "race" and antemortem stature are known? Several other corollary issues arose along the way. They are considered in situ.

MATERIALS AND METHODS

Sample

The "raw" data for this analysis consisted of summary statistics assembled from 55 samples of modern human populations documented in the literature. Many of these samples were collected between 1900 and 1950. As a result, the type of information included for each sample was highly variable. Only a few authors published raw femur length and stature for each individual, still fewer reported femur/stature ratios for each individual, and in only a few cases were statistics other than means reported for any sample. Since neither individual data nor measures of dispersion were available in very many cases, the present study is therefore based on the analysis of mean data from 55 samples that summarize 10,668 femur/stature pairs. The pertinent details about these samples are listed in Table 1.

In earlier studies of the femur/stature ratio (i.e., Feldesman and Lundy, 1988; Feldesman et al., 1990) samples were included as long as a femur/stature ratio could be obtained either by computing it from average femur lengths and average statures, or from a published value. As can be demonstrated easily, the ratio of average femur length to average stature is not computationally equivalent to the average femur/stature ratio. To maintain methodological consistency insofar as possible, we constrained the present study to include only samples for which average femur lengths and average statures were reported (or could be computed) from the primary literature. This allowed us to directly compute the ratio of average femur length to average stature for each sample included in the study. We imposed this restriction to insure that all femur/stature ratios were computed using comparable numerators and denominators. We did not use published values of the ratio unless we could verify that, in the absence of the necessary numerator and denominator, the calculation used the appropriate values. As a result, we eliminated seven samples that FKL used in the earlier study; however, we found additional "new" data, and our design permitted us to utilize "old" data more efficiently. Collectively this added 16 "new" population samples to those recycled from FKL, and yielded a total sample of 55 populations—a net increase of nine from 1990.

As stated earlier, our primary concern in this study was to determine whether we could detect statistically significant "racial" differences in the femur/stature ratio. In 1990, FKL conclusively demonstrated that gender differences in the ratio were not statistically significant and we, therefore, do not reconsider them here.

We followed the 1990 protocol in constructing "racial" categories into which the 55 samples could be placed. We again subdivided the 55 samples into three broad quasigeographic "races": "Whites," "Blacks," and "Asians." We recognized from the outset that this grouping was crude and did not capture the more substantial store of geographic variation characteristic of modern humans. However, the paucity of appropriate skeletal data mitigated against any finer-grained subdivisions, such as those based on latitude or other more specific aspects of geography including climate and thermoregulation. For most applications where these ratios might be employed (e.g., in paleoanthropology), we believe these subdivisions are adequate. In inspecting our classifications, few would argue with our assignment of specific samples to the "White" or "Black" groups. There is,

TABLE 1. Populations used in this study[1]

Population	"Race"	Original N	Femur length	Stature	Ratio
US Asian males military	A	67	44.25	167.84	26.36
US Asian males military	A	60	44.64	168.45	26.50
US Mexican males rt military	A	50	45.14	168.62	26.77
US Mexican males lt military	A	57	45.60	169.00	26.98
US Puerto Rican males rt military	A	40	44.76	166.40	26.90
US Puerto Rican males lt military	A	44	44.82	166.93	26.85
Mesoamerican (Indian) males	A	22	43.21	163.99	26.35
Mesoamerican (Indian) females	A	15	39.63	152.30	26.02
Han Chinese	A	100	42.96	161.06	26.67
East India Lucknow	A	40	43.39	160.62	27.01
Calcutta males	A	86	41.76	160.96	25.94
Calcutta females	A	56	38.72	148.73	26.03
Eskimo	A	3	39.93	155.23	25.72
US Black males rt military	B	343	48.22	173.59	27.78
US Black males lt military	B	338	48.39	173.74	27.85
US military Black males	B	54	48.34	172.11	28.08
US Black Terry males	B	360	47.42	172.73	27.46
American Black males	B	100	47.72	176.22	27.08
American Black females	B	100	43.96	164.06	26.80
US Black Terry females	B	177	43.71	160.89	27.17
US Black males	B	19	46.10	169.20	27.25
US Black females	B	16	43.80	160.60	27.27
South African Black males	B	175	44.77	162.93	27.48
South African Black females	B	122	42.33	154.12	27.47
Occidental Pygmy males	B	5	40.36	150.20	26.87
Occidental Pygmy females	B	3	37.61	144.16	26.09
Oriental Pygmy males	B	10	36.29	141.19	25.70
Oriental Pygmy females	B	5	35.31	132.96	26.56
German females	W	500	42.40	161.80	26.21
Finnish males	W	125	45.48	169.40	26.85
Finnish females	W	47	41.78	156.80	26.65
US White males rt military	W	2327	47.08	174.32	27.01
US White males lt military	W	2345	47.15	174.27	27.06
US White military	W	710	46.97	174.04	26.99
US White Terry collection	W	255	45.66	170.39	26.80
US White females Terry	W	63	42.96	160.68	26.74
Greek males	W	288	43.17	167.60	25.76
Greek female	W	126	41.17	156.68	26.28
European mix	W	135	44.32	169.08	26.21
European mix	W	140	46.22	170.38	27.13
French White males	W	94	43.91	163.99	26.78
French White females	W	82	40.35	152.20	26.51
American White males	W	100	45.33	172.96	26.21
American White females	W	100	42.24	160.96	26.24
French White males	W	50	44.52	165.00	26.98
French White females	W	50	40.86	152.30	26.83
Young White males	W	35	44.82	169.09	26.51
Middle young Whites	W	123	44.78	168.26	26.61
Middle-aged White males	W	113	44.87	167.70	26.76
Old White males	W	63	44.86	165.14	27.16
Young White females	W	17	42.46	158.96	26.71
Middle-aged White females	W	33	41.86	157.41	26.59
Late middle-aged White females	W	39	41.73	157.11	26.56
Old White females	W	41	41.83	154.70	27.04
Miscellaneous European males	W	200	46.27	170.46	27.14

[1] Data sources include Bach (1965), Dupertuis and Haddon (1951), Eliakis et al. (1966), Feldesman and Lundy (1988), Genoves (1967), Hrdlicka (1972), Marquer (1972), Mo (1983, 1984), Nat (1930), Olivier (1963), Olivier and Tissier (1975), Olivier et al. (1978), Pan (1924), Rösing (1983), Telkka (1950), and Trotter and Gleser (1952, 1958). "Race" abbreviations are A ("Asian"), B ("Black"), and W ("White").

however, a notable problem in assigning data from male and female US-resident Mexicans and Puerto Ricans reported by Trotter and Gleser (1958). FKL excluded these data in 1990. We included them here, fully recognizing the difficulty of doing so. According to our classification scheme, their only plausible assignment was to the "Asian" subgroup. Nonetheless, we recognized that both groups contain members who are "Latino"—relatively recent admixtures of European and indigenous populations (see Williams, 1994, for a discussion of the extent of genetic admixture for Mexican Americans). To deal with this

classification problem, we performed our analyses in two ways—first by treating these four samples as "Asians," and second by leaving them out of the analysis altogether. There were too few to treat them as a separate group, which was the preferred strategy. As our results will show, the impact of this admixture is quite apparent.[2]

Statistical analyses

Analysis of variance. We used analysis of variance (ANOVA) as the primary technique for addressing the question of whether the femur/stature ratios were significantly different in our three "racial" subdivisions. Our null hypothesis was that these three "races" sampled the same population. We used a nested, mixed-effects design (Neter et al., 1990). The assumed model was:

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \epsilon_{(ij)k}$$

where μ is the overall mean, τ_i is the fixed effect due to race, $\beta_{j(i)}$ is the random effect due to subgroup, nested within race, and $\epsilon_{(ij)k}$ is the individual error term. The index i runs from 1 to 3 ("White," "Black," "Asian") and the index j from 1 to 27 for "White," 1 to 15 for "Black," and from 1 to 13 for "Asian." The index k runs from 1 to the number within each subgroup (e.g., 500 for German females).

If the above design were balanced, then the expected mean squares for the factors would be:

Factor	EMS
τ_i	$bn \sum \frac{\tau_i^2}{a-1} + \sigma^2_{\beta(\tau)} + \sigma^2$
$\beta_{j(i)}$	$\sigma^2_{\beta(\tau)} + \sigma^2$
$\varepsilon_{(ij)k}$	σ^2.

The F-test for the main effect would then use the test statistic

$$\frac{MS_\tau}{MS\beta_{(\tau)}}.$$

Since the design is unbalanced, this statistic provides an approximate main-effects F-test. Because the main-effects test is the only one of interest, it is not necessary to calculate an estimate of σ^2. We followed the ANOVA with the Tukey multiple comparisons test, which computes all pairwise differences and identifies "race" pairs responsible for significant F-values.

After completing the ANOVA on the ratios, we performed an analysis of covariance (ANOCOVA) of femur length on stature to test whether the regression slopes were parallel and the regression intercepts were equal across all three groups.

The ANOVA and ANOCOVA were performed using SAS-PC's (SAS Institute Inc., 1989) "REG" and "GLM" routines, respectively.

Approximate randomization. We recognized from the beginning that our assignment of individuals to "treatment" groups (i.e., "races") was not random. Consequently, we felt that the variance ratio, which is used to calculate the F-statistic, might not necessarily follow the F-distribution. To overcome this problem we used Monte Carlo simulation to estimate the distribution of our variance ratio under nonparametric conditions. The preferred simulation technique was an exact randomization test (Noreen, 1988; Sokal and Rohlf, 1995; Westfall and Young, 1993) in which all possible random partitions of 27, 15, and 13 individuals selected from a sample of 55 are considered. While this would have yielded the exact distribution, it would have necessitated computing variance ratios from approximately 1.43×10^{23} replicates, and thus was impractical.

An alternative was the "bootstrap" (Efron and Tibshirani, 1993), which involves resampling (sampling with replacement) the original data set using its original partitions to guide the construction of the derived samples. In other words, it would have required resampling the "Whites," resampling the "Blacks," and resampling the "Asians" before any further analysis of the data. Unfortunately, this technique assumes the original partitioning of the data is valid a priori. Since we had no way of knowing for certain that this was the case, and indeed had rea-

[2]Our "Asian" group also includes populations clearly identified as "Mexican." Genoves (1967) collected data on both "mestizo" and "Indian" specimens. We limited the data used here to only those specimens that Genoves classed on serological grounds as "Indian." This avoids this issue of admixture in these groups.

sons to believe it might not be, we rejected the "bootstrap."

Instead we used a third technique called approximate randomization (Noreen, 1988) or sampled randomization (Sokal and Rohlf, 1995). Our interest was in building an approximate distribution for our specific data set, and to determine the likelihood of obtaining *worse* results than our parametric F-value by chance alone. To do this, we took a random sample of 55 without replacement (wr) from the original 55 populations. The first 27 members sampled were assigned to the "White" group (whether they were "White" or not), the next 15 were assigned to the "Black" group, and the last 13 were assigned to the "Asian" group. We then computed an ANOVA on this sampling replicate and recorded the F-value (variance ratio). We generated 10,000 samples following this procedure. By examining the ordered table of 10,000 variance ratios (F-values?), we were able to establish the empirical probability of obtaining a *larger* (i.e., worse) variance ratio than we obtained in the original ANOVA. In so doing, we were able to assess the validity of the parametric F-statistic. This approach appears to be a straightforward approximate randomization test; however, there is a bootstrap component that may not be evident to the reader. In a true randomization experiment, only unique sample partitions would be used. We did not determine whether we had duplicate sample partitions. Thus, while we constructed our individual sample partitions of 27, 15, and 13 *without* replacement and ensured that no single data point appeared in any replicate more than once, we extracted our 10,000 samples of 55 from the universe of 1.43×10^{23} *with* replacement. We made no effort to determine the frequency of replicate redundancy since, with this number of replicates, bootstrapping is quite robust. We coded the approximate randomization test using the macro language, and built-in sampling, simulation, and conventional statistical primitives of STATA Version 4.0 (Stata Corporation, 1995).

"Misclassification" testing. We performed two "misclassification" tests—*k*-means cluster analysis (Hartigan, 1975) and discriminant analysis (Morrison, 1990)—first using only the group ratios, and then only the ratio components, femur length and stature. We used SAS-PC's "DISCRIM" and "FASTCLUS" (SAS Institute, 1989) to perform these analyses. Each test was done in two ways, first by including the US Mexicans and US Puerto Ricans, and then by excluding them.

We performed two separate *k*-means analyses, each with the number of clusters set at three. In the first case we used the group ratios as the grouping variable; in the second instance we used both femur length and stature. We then compared the actual groups with the assigned groups based on the cluster analysis and assembled misclassification statistics from these data.

We also used discriminant function analysis for assessing misclassification frequencies. Our protocol was identical to that used above in the *k*-means analysis. We used the same data configuration (i.e., with and without Mexicans/Puerto Ricans) and the same variable suite (i.e., ratios, femur length, and stature).

Validation testing. To test the validity of our ratios, we assembled six samples, representing all three "racial" groups, where we knew individual femur lengths and individual statures a priori. The total n for these was 798. These included the Vietnam veterans sample of 50 White males described in FKL, a sample of 297 male and female Black (Bantu) South Africans originally described by Lundy (1984), a sample of 142 East Indians described by Nat (1930), a sample of 40 East Indians described by Pan (1924), a sample taken by one of our students of 250 Blacks and Whites from the Terry Collection, and a collection of 19 autopsy specimens from the forensic collection at the Maxwell Museum at the University of New Mexico. We predicted statures for each individual using their femur lengths and (1) the appropriate "race"-specific ratio from a "three-race" model; (2) the appropriate "race"-specific ratio from a "two-race" model; and (3) the generic ratio.

We used three procedures to test validity. In the first, we compared predicted statures (using each of the three predictor formulae described above) for every individual. We then calculated the mean absolute deviation

$$\text{MAD} = \frac{\sum|(\text{computed stature} - \text{actual stature})|}{n},$$

which is the average of the absolute deviations. The formula with the *lowest* MAD was judged the best estimator. Second, we evaluated the mean squared error

$$\text{MSE} = \frac{\sum(\text{computed stature} - \text{actual stature})^2}{n}$$

for each of the predictor formulae. The best estimator had the *lowest* MSE.

Since the MAD is the average absolute error and the MSE is the average squared error, there are situations in which they can yield contradictory results. Hence it is useful to consider other estimation criteria as well. One such criterion is Pitman's measure of closeness (PMC). It has been shown to be invariant over different types of "loss" functions (absolute errors vs. squared errors, for example). Thus, if MAD and MSE disagree, PMC can be used to settle the dispute. Unlike the MAD and the MSE, which are concerned with the average loss and consider only the marginal distributions of the estimators, PMC is concerned with the simultaneous behavior of several estimators and takes into account their joint distribution. The PMC of one estimator relative to another is the probability that the first estimator is closer than the second to the parameter (Keating et al., 1993). Using MAD or MSE, one estimator is preferred over another if the first has a smaller average loss than the second. Using PMC, the first estimator is preferred if it has a higher probability of producing the estimate closest to the true value of the parameter. We used two variants of PMC—the simultaneous and the pairwise PMC. The Pitman algorithm and its computational details are described in Keating et al. (1993). We calculated the PMC statistic using Microsoft Excel Version 5.0 for Windows.

Since the three criteria (MAD, MSE, PMC) do not always agree, we felt it important to look at all three before making final judgment on the superiority of any of the particular stature-estimating techniques used in this study.

RESULTS

Table 2 details the descriptive statistics for each of the three groups in our sample, while Figure 1 depicts box plots of the distributions in each group. As can be seen, regardless of sample composition, these basic statistics confirm that "Blacks" have the highest femur/stature ratio, while "Asians" have the lowest ratio. The "generic" or combined ratio, which is obtained by combining all 55 (51) observations into a single group, is 26.75 (26.77). This is nearly identical to the value 26.74 reported in FKL.

Tables 3 and 4 summarize the results of the two ANOVA calculations done to assess the significance of the mean differences documented Table 2. Table 3 reports the ANOVA for the study including the US Mexican and Puerto Rican samples as "Asians," while Table 4 figures the results with those four samples omitted from the analysis. Table 3 clearly demonstrates that when all samples are included in the analysis, the three mean "racial" femur/stature ratios differ significantly from one another ($F = 7.50$; $P = 0.0014$; df = 2,52). Post hoc comparisons using the Tukey HSD procedure (Lindman, 1992) show that the significant ANOVA results from differences between "Black" and "White" means, and between "Black" and "Asian" means ($P = 0.012$ and 0.002, respectively). The "White" and "Asian" means are not significantly different from one another ($P = 0.39$).[3]

Table 4 details the results of the ANOVA

[3]An anonymous reviewer suggested that one reading of Williams (1994) might have the US Mexican and US Puerto Rican populations more "europeanized" than "asianized." Consequently, the reviewer suggested we might consider these admixed populations in with our "White" group rather than with the "Asians." Although every treatment of these US Mexican and US Puerto Rican samples has placed them with "Asians" whenever a triracial classification scheme has been used, we were nonetheless curious to see what would happen in our analysis if these four samples were shifted to the "White" group from the "Asian" group. The analysis of variance results were highly significant ($F = 10.05$, $P = 0.0002$); however, the post hoc tests again failed to distinguish between "Asian" and "White" populations. While these groups were more distinct than when the admixed populations were considered "Asians," this did not evidence itself in a significant post hoc comparison. On the other hand, the shift exaggerated the "Asian"–"Black" and "White"–"Black" differences even more.

TABLE 2. Femur/stature ratio means and standard deviations for three "racial" subdivisions[1]

"Race"	Observations	Mean	Std Dev	Observations	Mean	Std Dev
Asian	13	26.471	0.435	9	26.494	0.452
Black	15	27.126	0.647	15	27.126	0.647
White	27	26.678	0.354	27	26.678	0.354
Combined	55	26.75		51	26.77	

[1] The "Asian" subgroup shows two different sample sizes because US Mexicans and US Puerto Ricans are included in column 2 and excluded in column 5.

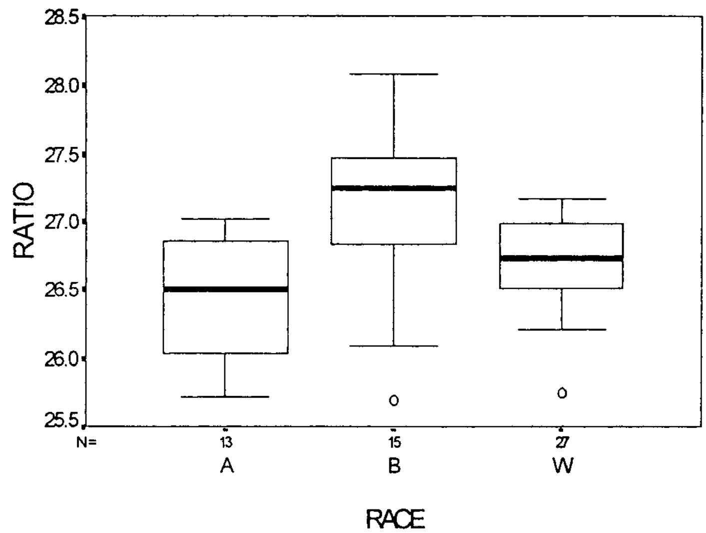

Fig. 1. Box and whisker plot of the femur/stature ratio for the three "racial" groups defined in this analysis. Boxes cover the 25th to 75th percentiles of the distribution. Horizontal heavy lines across each box represent the median of each distribution; the tails extend to the fifth and 95th percentiles while the small circles represent outliers that fall below the fifth percentile.

TABLE 3. Analysis of variance table for three-group "racial" analysis (US Mexicans and Puerto Ricans included)

Source	SS	df	MS	F	Prob > F
Between groups	3.28058	2	1.64029	7.50	0.0014
Within groups	11.3768	52	0.218785		
Total	14.6574	54	0.271433		

TABLE 4. Analysis of variance table for three "racial" groups (US Mexicans and Puerto Ricans excluded)

Source	SS	df	MS	F	Prob > F
Between groups	4.1556	2	2.0778	9.58	0.0003
Within groups	10.4131	48	0.21694		
Total	14.5687	50	0.29137		

excluding the US Mexican and Puerto Ricans from the analysis. This ANOVA, like the first, is highly significant ($F = 9.58$; $P = 0.0003$; df = 2,48). In fact, by merely excluding these hard-to-classify admixtures there is nearly a fivefold *decrease* in the probability that these results are due to chance. The Tukey test continues to attribute the significant F to differences between "Blacks" and "Whites" ($P = 0.012$), and between "Blacks" and "Asians" ($P < 0.0005$), while the "White"–"Asian" difference increases ($P = 0.09$), but remains insignificant. However, it is worth noting that removing the populations subject to "White"–"Asian" admixture decreases the probability that differences between these two "racial" groups are due to chance alone from 0.39 to 0.09.

We conducted an analysis of covariance (ANOCOVA) with stature as the dependent variable and femur length as the covariate. When all 55 samples were included in the analysis, we found significant differences in both slopes and intercepts ($P < 0.001$) be-

tween "Asian" and "Blacks," and "Whites" and "Blacks," while the "Asian"–"White" pairing produced nonsignificant slopes ($P = 0.3999$) and intercepts ($P = 0.3995$).

When we removed the admixed populations from the analysis, slopes and intercepts between "White"–"Black" and "Asian"–"Black" still were significantly different at $P < 0.001$, while the "Asian"–"White" differences continued to be nonsignificant. As in the Tukey test, however, the probabilities that the "Asian"–"White" slope and intercept differences were due to chance alone *declined* significantly (slope $P = 0.1687$; intercept $P = 0.1991$).

We subjected both sample configurations to an approximate randomization test. Using the full sample, only 16 sample permutations (out of 10,000) yielded variance ratios larger than the recorded 7.50 for the parametric ANOVA. This means that the approximate empirical probability level associated with ANOVAs constructed from 10,000 random samples of 27, 15, 13 taken from the original 55 is $P = 0.0016$. This makes it highly unlikely that our three "racial" groups could have been sampled from a single population by chance alone. This compares with the parametric probability based on the F-distribution of $P = 0.0014$, and suggests that the F-distribution is appropriate for our data set.

We repeated this simulation deleting the US Mexican and US Puerto Rican samples. We found only four sample replicates with variance ratios larger (i.e., worse) than the observed parametric F. Therefore, the empirical probability of finding an F-value larger than that yielded by our original sample was only $P = 0.0004$. Again, this compares quite favorably with the parametric probability from the F-distribution ($P = 0.0003$), and suggests that the ANOVA was appropriate and valid.

Table 5 summarizes the results of the different classification functions constructed from different combinations of variables in the data set. Regardless of the technique (*k*-means or discriminant analysis), variables (ratio or ratio components), or sample composition (including US Mexicans and Puerto Ricans or excluding them), the message is the same. The groups are not very coherent, with individual samples misassigned 33% of the time in the best circumstance, and 57% of the time in the worst case. As Table 5 clearly shows, the misclassification rates vary widely across groups depending on the grouping technique and variables included.

Finally, we focused on the validity of the ratios in predicting stature from femur length, when we knew stature a priori. We were primarily interested in how accurately each ratio predicted actual stature. Three different validation configurations emerged from the ANOVAs, the post hoc tests, and the 1990 FKL results. We used the six data sets sampling all three "races" for validation. First, we predicted individual statures using the appropriate "race"-specific femur/stature ratio ("three-race" model). Second, since we found no significant difference between the "White"–"Asian" pair, we computed "race"-specific ratios by combining "Whites" and "Asians" into a single group, while calculating a separate ratio for "Blacks" ("two-race" model). Finally, we predicted stature using the generic ratio as advocated in FKL.

On the basis of the mean absolute deviation criterion, the "three-race" model ratios had the lowest MAD value and were thus the "best" predictor of stature when "race" was known a priori (MAD = 3.880 cm). In the "two-race" configuration, the MAD was only fractionally less accurate than the "three-race" model (MAD = 3.897 cm). Finally, the generic ratio yielded the worst results (MAD = 4.426 cm).

The mean squared error criterion yields results identical to the MAD. The "three-race" ratios yield the lowest MSE (25.281), the "two-race" ratios are only slightly larger (MSE = 25.834), while the generic ratio does the poorest (MSE = 32.651).

As noted in the Materials and Methods section, the aforementioned validators consider only the marginal distributions of the estimators. The Pitman measure of closeness (PMC) provides a different perspective on the overall performance of these three estimators. When the criterion is applied to the estimators two at a time (i.e., pairwise), the "three-race" ratios yielded better results (closer fit between actual and predicted statures) than the "two-race" ratios 52% of the time. Both the "three-race" ratios and the "two-race" ratios outperformed the generic ratio 61% of the time. The PMC performance hierarchy is identical to results from the

TABLE 5. Misclassification rates for k*-means cluster analysis and discriminant analysis*

Analysis	Variables	Mex/PR	Percent misassigned			
			White	Black	Asian	Total
k-means	Ratio	Yes	22	60	54	40
k-means	Ratio	No	22	60	33	35
k-means	Fem,stat	Yes	56	73	38	56
k-means	Fem,stat	No	52	92	33	57
Discrim	Ratio	Yes	63	33	46	51
Discrim	Ratio	No	52	33	33	43
Discrim	Fem,stat	Yes	44	27	46	40
Discrim	Fem,stat	No	55	27	33	33

TABLE 6. Pitman measure of closeness (PMC) matrix based on simultaneous analysis of all three estimators[1]

	"Three-race" ratio	"Two-race" ratio	Generic ratio
Best	0.318	0.304	0.378
Middle	0.492	0.487	0.021
Worst	0.190	0.209	0.601

[1] Numbers in each column represent proportion of cases in which the specific estimator yielded the closest approximation to actual stature, the next closest approximation to actual stature, and the worst approximation to actual stature. Note that the column totals and the row totals all equal 1.

MAD and MSE validation hierarchy (i.e., "three-race" > "two-race" > generic).

The simultaneous (three at a time) application of the PMC criterion yields slightly different (and possibly contradictory) results. Table 6 details the Pitman matrix for simultaneous comparisons. This table shows that each of the three estimators gives a "best" performance (closest match between actual and predicted statures) more than 30% of the time. Closer inspection of the "best" row of this table shows, perhaps surprisingly, that the generic ratio is the "best" single predictor more frequently than the other two. At the same time, the category "middle" (neither "best" results nor "worst" results) follows according to expectations with the performance hierarchy "three-race" > "two-race" > generic. Finally, the "worst" category (largest difference between actual and predicted statures) is "won" by the generic ratio. In other words, the generic ratio yields the "worst" outcome far more often than any of the other equations. Thus, we have the paradoxical situation in which the PMC simultaneous comparison tells us that the generic ratio performs "best" more often than either of the "race"-differentiated ratios, but also performs "worst" more often than the others.

DISCUSSION

In 1990 FKL claimed that the generic femur/stature ratio yielded reliable estimates of living stature, and should be used in cases where neither gender nor "race" could be determined confidently. In that analysis, they also considered whether subdividing the samples into broadly defined geographic "races" would produce different and better point estimates of individual statures. Their preliminary results suggested that the mean femur/stature ratios of three crudely defined "races" were statistically different. At the same time, FKL's testing showed that even when "race" and gender were known a priori that information did not offer significant improvements in stature-estimating accuracy. In some instances the particularist estimates proved to be less accurate than those yielded by the more generic approach. FKL did not preclude the possibility that with larger and more well defined samples these "race"-specific ratios might prove to be better stature estimators.

In formulating the present study, we were interested in determining whether more samples, improved categorization, and a better research design could resolve the questions of the significance of differences among the three quasigeographical "races" in femur/stature ratios and whether these differences translate into significantly higher prediction accuracy than the generic ratio proposed in 1990.

The present study unequivocally resolves the first matter. Our ANOVA unambiguously confirms that there are statistically significant differences in the mean femur/stature ratios of the three geographic "races" constructed for this analysis. At the same

time, multiple comparison tests, done to resolve the significant F-tests from the ANOVAs, make it clear that the "Black" femur/stature ratio is the outlier; neither "White" nor "Asian" ratios are statistically distinct from one another. These results arise whether we include the admixed US Mexican and Puerto Rican populations as "Asians" or exclude them from the analysis. When they are included in the analysis, the difference between "Whites" and "Asians" decreases. This result is not unexpected given that we know that both populations are European–Indigenous hybrids (with a small African component) and are intermediate in some important phenotypic (and genotypic) characters between these two target populations. Williams (1994) concludes, for example,that 68% of the alleles in the Mexican–American gene pools thus far sampled come from Europeans. This demonstrates the extent of the admixture, and also provides an underlying genetic basis for our expectation that including them as "Asians" would bias the "Asian" group toward the European "Whites." When we exclude these obvious hybrids from our analysis, the "White" and "Asian" ratios become more distinct, but not enough to be considered statistically significant.

The analysis of covariance of femur length on stature corroborates the ANOVA and multiple comparison tests. This also localizes the differences and indicates that the "Black" slope and intercept are significantly different from the "White" and "Asian" slopes and intercepts, which are themselves not distinct from one another.

The aforementioned results leave no doubt that the "Black" sample presents a significantly different and distinct femur/stature ratio. At the same time, the results fail to justify any claim that "Whites" and "Asians" evidence any proportional differences in their femur lengths relative to their statures.

An obvious question arises from our analysis: how reasonable are the groupings? We acknowledge freely that the groupings are only loosely based on geographic and genetic propinquity. We would, of course, challenge any worker faced with the same set of data to figure out how they might be organized differently (other than to pool them) to enable any sort of valid statistical testing of "racial," ethnic, or population differences in the ratio. Nevertheless, we were concerned with the reliability and stability of our groupings. The purpose of performing classification testing was to provide a basis for assessing confidence in the way we grouped the data.

The results of our classification studies make it painfully clear that it is extremely difficult to form discrete, coherent, and nonoverlapping groups. This comes as no surprise, given our knowledge of the extent to which all contemporary human populations have been exposed to admixture. We were, however, surprised at just how much overlap existed, and how this overlap came to affect the way groups were formed statistically, and influenced the assignment of known individuals to their "correct" group. In short, the classification tests yielded "correct" assignments no less than 43% of the time and no more than 67% of the time. These results are not much better than one could expect to obtain by random assignment alone, despite extensive geographic and cultural provenience data. These results would hardly surprise forensic anthropologists (e.g., Bass, 1987; St. Hoyme and Işçan, 1988), who have long maintained that the postcranial skeleton is not diagnostic of "race."

When FKL proposed using the generic femur/stature ratio in 1990, they envisioned its use in estimating stature from femora for which neither "race" nor gender could be ascertained with any confidence. The intended application then, as it remains now, was in paleoanthropology where there is never any substantive basis for assigning "race" in order to choose a modern "race"- and gender-specific regression equation. FKL demonstrated, and this study reaffirms, that a reasonably good estimate of stature can be obtained simply by knowing that across the human-populations sampled, femur length constituted 26.74 (26.75%) of stature, and that no assumptions about "race," ethnicity, or gender were required.

Despite the congruence of the present findings and those from FKL, we felt it important to continue testing this ratio against alternative estimators suggested by the

ANOVA. We selected three different ratio forms for validation, and used four different validation statistics for our tests. As the results we reported showed, the "three-race" femur/stature ratio barely outperformed the "two-race" ratio using the mean absolute deviation and mean squared error criteria, but both ratios yielded better estimates than the generic ratio. More significantly, however, the magnitude of the deviation evidenced by the MAD criterion was 3.8 cm for the "race"-separated ratios, but only 4.4 cm for the generic ratio—a difference of only 0.6 cm between the ratios.

The Pitman pairwise criterion also suggested that the "race"-specific ratio was a better stature estimator than the other two ratios. The Pitman simultaneous criterion confused matters in reporting that when all three estimators are considered *at the same time*, it is the generic ratio, not the "race"-specific ratio, that performs best most often. On the other hand, however, the generic ratio also performed worse more often than the other ratios.

CONCLUSIONS

The results of the present investigation offer some resolution to problems left unresolved at the end of FKL's study. There is no doubt that different geographic "races" present different body proportions, and these body proportions translate into different femur/stature ratios. Every result presented here corroborates that point, and strongly suggests that "three-race" (or "two-race") ratios should be used wherever possible.

Unfortunately, while these results are clear, two fundamental questions arise: (1) when is it *possible and appropriate* to use "race"-specific ratios, and (2) what penalty in accuracy is paid if one eschews the assumptions of the "race"-specific ratios (or population- and gender-specific regression equations) and uses the generic femur/stature ratio instead?

The answer to the second question is easy. Our MAD results suggest that using the generic ratio will cost about 0.6 cm of accuracy compared with using the "racially correct" ratio. This simple and obvious answer conceals another significant problem. Our classification results indicate that selecting the "correct" equation is more difficult than it might appear to be. Our studies suggest that known geographic and cultural provenience is not particularly diagnostic of geographic "race" when femur length, stature, or the ratio between them are the only variables. There is so much morphologic overlap between groups that are geographically distinct that we cannot accurately discriminate "races" even when we have access to the means of femur length and stature for every group composing the geographic races. Our results might improve if we had access to *all* the underlying raw data, but the forensic literature suggests otherwise. On the other hand, if we had those data we could construct individual regression equations for every population, and use the "correct" equation for every specimen known to come from that group. However, this still would do us no good in estimating stature of specimens known not to have come from any group for which an estimating equation exists. Without those data, we are left with the present approach in which one can expect to be correct, on average, no more than about 50% of the time even with all of the important provenience information.

Given this, how can one decide which equation to use, and more significantly, what penalty does one pay for choosing a specific equation incorrectly? To answer this, we calculated stature for each specimen in the validation set using *each* of the three "race"-specific ratios. We then computed the MAD and MSE for each predictor for each specimen. By averaging the absolute deviations for all three predictors, we were able to capture the entire sample space and get *expected* values for MAD and MSE. In this approach one correctly assigns "race" by chance alone only one-third of the time. This technique provides us with the expected value for the MAD of 4.632 cm based on all possible racial assignments (correct and incorrect), and an expected value for the MSE of 35.388. Both of these values exceed the equivalent measure for the generic ratio. These results are sobering. If one knows nothing a priori about

TABLE 7. Regression equations for predicting stature from femur length, based on data from Table 1

"Three-race" regression equations:	
Asian	= 40.167154 + 2.841734 × Femur
Black	= 30.285687 + 2.986895 × Femur
White	= 21.676678 + 3.254227 × Femur
"Two-race" regression equations:	
Black	= 30.285687 + 2.986895 × Femur
White–Asian	= 29.737448 + 3.074994 × Femur
Generic regression equation:	
Generic	= 31.263362 + 3.019390 × Femur

TABLE 8. Regression and ratio analyis results using validation suite[1]

	MAD	MSE
"Three-race" regression	3.266	17.888
"Two-race" regression	3.259	18.086
Generic regression	3.787	23.126
Regression random assignment	3.992	25.607
"Three-race" ratio	3.880	25.281
"Two-race" ratio	3.897	25.834
Generic ratio	4.426	32.651
Ratio random assignment	4.632	35.388

[1] Four different analyses were run using different configurations of the data. The random assignment regressions (ratios) were run by computing stature estimates for every specimen in the validation set using the "race"-specific regression (ratio). Each specimen was classified in turn as "White," "Black," and "Asian." The MAD and MSE statistics are the averages of the within-specimen averages.

the "racial" background of the specimen, and guesses "race" incorrectly, there will be a larger estimation error than would be obtained by using the generic ratio. Even if one has reason to believe the assignment is correct, and it turns out not to be, the error will again be greater than by using the generic ratio. The results of this simple exercise confirm that only if one knows in advance that the "racial" assignment is correct can one choose a "race"-specific equation with impunity. In that case, and only in that case, will the estimating errors be smaller than those obtained by using the generic ratio, but only by about 0.6 cm.

In general, regression analysis is regarded by most statisticians (e.g., Moore and McCabe, 1993) as the fundamental technique for predicting the value of a dependent variable from an independent variable. Thus, it is not surprising that regression equations have found widespread use in the stature-estimating literature. As a result, we wanted to compare our ratio estimators of stature with some regression estimators. Therefore, we computed our own suite of regression equations based on the same data set used to calculate the ratios, and then predicted statures for each of the cases in our validation suite. This enabled us to directly compare our ratio estimators with regression estimators based on the same data set. Table 7 summarizes the regression equations based on the three different configurations of our data set: (1) "three-race" regression equations; (2) "two-race" ("Whites" and "Asians" consolidated) regression; and (3) no "races" ("generic" regression).

The results of the regression analysis mirror those from the ratio study. Table 8 compares the MAD and MSE results for the four different configurations of the validation data suite for both regression and ratio studies. Both the MAD and MSE results for the regressions are lower than their ratio counterparts. Again, these results confirm that reasonable results can be obtained with a "race"-specific regression equation *only* if one is confident in the racial assignment. If the racial assignment is incorrect, as the "random assignment" results show, there is a substantial estimation penalty.

The central factor motivating FKL was the difficulty of estimating stature from femora of fossil hominids for which neither gender nor "race" was (or could be) known. This difficulty is the paleoanthropological analog of the classification problem described above and for which there is no remedy. One occasionally gets corroborating skeletal material that confirms gender, but we would argue that admixture makes geographic provenience uninformative even in contemporary "racial" classification. At an earlier temporal provenience, admixture may have been less of a problem, but the modern equations, deeply rooted in gender and population specificity, constrained by size ranges, and confounded by admixture, are no antidote. Their use makes for poor stature estimates in fossil and more recent prehistoric hominids, as numerous recent studies, including FKL, demonstrated. That conclusion, coupled with FKL's advocacy of the simpler and more as-

sumption-free femur/stature ratio, has prompted a number of workers to abandon regression-based stature-estimating equations in favor of our ratio for hominid fossils (see, e.g., McHenry, 1991).

Both the present results and those from FKL strongly suggest that caution should be exercised when using forensic regression equations to predict stature of individuals for whom gender and race attribution is uncertain. We have demonstrated here that when only "race" is considered, and the stature estimates compared using ratios and regression equations built from the same data set, the regression equations offer a "best-case" estimate improvement of about 0.4–0.6 cm in matched circumstances ("three-race" ratio vs. "three-race" regression equations, generic ratio vs. generic regression, etc). However, when the comparison is between the "best-case" regression ("three-race") and the "worst-case" ratio (generic), the difference in accuracy is only 1.2 cm. Given the clear difficulty with assigning femora to "racial" categories without other corroborating evidence—even in modern specimens—the most pragmatic strategy would be to use a generic regression or a generic ratio, as opposed to guessing. The cost of using the generic ratio instead of the generic regression is only about 0.6 cm.

In the final analysis, we favor any technique that yields relatively reliable answers with a minimum of essential information. Our study of estimating errors resulting from the various approaches to predicting stature favors a generic approach over any approach that is bound to assumptions of "race" or gender. The generic femur/stature ratio has been well tested and requires workers to pay a relatively small price (<1.2 cm) for the luxury of making no assumptions about gender and "race." The advantages of the generic approach become even more apparent when it is weighed against the estimating penalties that arise by choosing an incorrectly specified regression equation. A "race"-specific ratio offers little protection against a wrong specification, for here, too, we have shown that an improper choice of "races" yields significantly worse stature estimates than the generic ratio.

While we favor the generic ratio on the basis of its simplicity and its solid performance, we would be remiss if we did not note our incidental finding that a generic regression of femur length on stature yields even more accurate results than the generic ratio. We have not done any rigorous testing of this equation, and do not yet have a clear feeling for its performance over a wide range of data. But for those workers interested in predicting the stature of fossil hominids using this equation, we have provided both the equation and the raw data used to generate it. We do not recommend its uncritical use until we, or someone else, can subject it to the analysis it requires. In the meantime, the generic femur/stature ratio has been thoroughly tested and proven to be a very reasonable estimator of stature. We recommend its continued use for fossil hominids until it can be shown that better estimators, equally or more free of assumptions, exist across the full range of human statures. At the same time, the findings presented here and in FKL make it clear that paleoanthropologists should no longer use modern, population- and gender-specific regression equations to predict stature in fossil hominids.

ACKNOWLEDGMENTS

We gratefully acknowledge the following researchers for making validation samples available for our use: Dr. Ellis Kerley for the Vietnam veterans sample, Dr. Stanley Rhine for forensic material from the Maxwell Museum (University of New Mexico), Ms. Tess Friedenburg for collecting femur lengths and stature data from the Terry Collection (Smithsonian Institute), and Dr. John Lundy for data from the Raymond Dart Collection (University of the Witwatersrand). Dr. Sharon Carstens translated Chinese literature and Mr. Vladimir Alexe'ev assisted in the translation of some Russian material. Various drafts of this manuscript were read and critically reviewed by Dr. Susan Wolf, Dr. John Lundy, Mr. Geoff Kleckner, and Ms. Karla Schilling, whose comments measurably improved the finished version. We also thank the anonymous reviewers for their helpful suggestions. We are grateful to the Applied Statistics Colloquium at Portland State University for making it possible to

collaborate on this project. M.R.F. owes a special debt of gratitude to Dr. Chris Ruff, whose persistent and thoughtful criticism of FKL, and subsequent work by Feldesman (1992a,b), inspired the present inquiry.

LITERATURE CITED

Allbrook D (1961) The estimation of stature in British and East African males. J. Forensic Med. *8:*15–28.

Bach H (1965) Zur Berechnung der Korperhohe aus den langen Gliedmassenknochen weiblicher Skelette. Anthropol. Anz. *29:*12–21.

Bass WM (1987) Forensic anthropology: The American experience. In A Boddington, AN Garlands, and RC Janaway (eds.): Death, Decay, and Reconstruction. Manchester, UK: Manchester University Press, pp. 224–239.

Dupertuis CW, and Haddon JA Jr (1951) On the reconstruction of stature from long bones. Am. J. Phys. Anthropol. *9:*15–51.

Efron B, and Tibshirani RJ (1993) An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability Number 57. London: Champman and Hall.

Eliakis C, Eliakis CE, and Iordanidis P (1966) Sur la determination de la taille d'apres la mensuration des os longs. Ann. Med. Leg. *46:*403–421.

Feldesman MR (1992a) The femur/stature ratio and estimates of stature in children. Am. J. Phys. Anthropol. *87:*437–459.

Feldesman MR (1992b) The bootstrap and race-specific estimates of the femur/stature ratio. Am. J. Phys. Anthropol. Supplement *14:*74.

Feldesman MR, and Lundy JK (1988) Stature estimates for some African Plio-Pleistocene fossil hominids. J. Hum. Evol. *17:*583–596.

Feldesman MR, Kleckner JG, and Lundy JK (1990) The femur/stature ratio and estimates of stature in mid- and late-Pleistocene fossil hominids. Am. J. Phys. Anthropol. *83:*359–372.

Formicola V (1993) Stature reconstruction from long bones in ancient population samples: An approach to the problem of its reliability. Am. J. Phys. Anthropol. *90:*351–358.

Genoves S (1967) Proportionality of the long bones and their relation to stature among Mesoamericans. Am. J. Phys. Anthropol. *26:*67–78.

Hartigan JA (1975) Clustering Algorithms. New York: John Wiley.

Holland TD (1992) Estimation of adult stature from fragmentary tibias. J. Forensic Sci. *37:*1223–1229.

Holland TD (1995) Brief communication: Estimation of adult stature from the talus and calcaneus. Am. J. Phys. Anthropol. *96:*315–320.

Hrdlicka AS (1972) Practical Anthropometry (original 1939). Philadelphia: Wistar Institute Press.

Jantz RL (1992) Modification of the Trotter and Gleser female stature estimation formulae. J. Forensic Sci. *37:*1230–1235.

Keating JP, Mason RL, and Sen PK (1993) Pitman's Measure of Closeness: A Comparison of Statistical Estimators. Philadelphia: Society for Industrial and Applied Mathematics.

Leakey REF, and Walker AC (1985) *Homo erectus* unearthed. Natl. Geogr. Mag. *168:*624–629.

Lindman HR (1992) Analysis of Variance in Experimental Design. New York: Springer-Verlag.

Lundy JK (1984) Selected Aspects of Metrical and Morphological Infracranial Skeletal Variation in the South African Negro. Ph.D. Thesis, University of the Witwatersrand, Johannesburg, Republic of South Africa.

Lundy JK, and Feldesman MR (1987) Revised equations for estimating living stature from the long bones of the South African Negro. S. Afr. J. Sci. *83:*54–55.

Marquer P (1972) Nouvelle contribution a l'etude du squelette des pygmees occidentaux du centre africain compare a celui des pygmees orientaux. Mem. Mus. Natl. Hist. Nat. Ser. A *LXXII:*1–122.

McHenry HM (1974) How large were the Australopithecines? Am. J. Phys. Anthropol. *40:*329–340.

McHenry HM (1991) Femoral lengths and stature in Plio-Pleistocene hominids. Am. J. Phys. Anthropol. *85:*149–158.

Mo S (1983) Estimation of stature by long bones of Chinese male adults in South China. Acta Anthropol. Sinica *2:*80–85.

Mo S (1984) Corrigenda to 1983 paper. Acta Anthropol. Sinica *3:*295–296.

Moore DS, and McCabe GP (1993) An Introduction to the Practice of Statistics. Second ed. New York: W.H. Freeman.

Morrison DF (1990) Multivariate Statistical Methods. Third ed. New York: McGraw-Hill.

Nat BS (1930) Estimation of stature from long bones in Indians of the United Provinces: A medico-legal inquiry in anthropometry. Indian Med. Res. *18:* 1245–1253.

Neter J, Wasserman W, and Kutner MH (1990) Applied Linear Statistical Models. Third ed. Burr Ridge, IL: Richard Irwin.

Noreen EW (1988) Computer Intensive Methods for Testing Hypotheses. New York: John Wiley.

Olivier G, and Tissier H (1975) Determination de la stature et de la capacite cranienne. Bull. Mem. Soc. Anthropol. Paris Ser: 13 *2:*1–11.

Olivier G, Aaron C, Fully G, and Tissier G (1978) New estimations of stature and cranial capacity in modern man. J. Hum. Evol. *7:*513–518.

Pan N (1924) Length of long bones and their proportion to body height in Hindus. J. Anat. *58:*374–378.

Pearson K (1899) Mathematical contributions to the theory of evolution. V. On the reconstruction of stature of prehistoric races. Philos. Trans. R. Soc. Lond. [Biol] *192:*169–245.

Peterson HC (1992) Multivariate Studies of Prehistoric Human Skeletal Remains: Mainly of Northern European Mesolithic Populations. Unpublished PhD Thesis, University of Aarhus (Aarhus, Denmark) and Universite de Bordeaux (Talence, France).

Reed CA, and Falk D (1977) The stature and weight of Sterkfontein 14, a gracile australopithecine from Transvaal, as determined from the innominate bone. Fieldiana Geol. *33:*423–440.

Rösing FW (1983). Stature estimation in Hindus. Homo *34:*168–171.

Ruff CB (1991) Climate, body size, and body shape in hominid evolution. J. Hum. Evol. *21:*81–105.

Ruff CB (1993) Climatic adaptation and hominid evolution: The thermoregulatory imperative. Evol. Anthropol. *2*:53–60.

Ruff CB (1994) Morphological adaptation to climate in modern and fossil hominids. Yearb Phys Anthropol. *37*:65–107.

Ruff CB, and Walker AC (1993) Body size and body shape. In AC Walker and RE Leakey (eds.): The Nariokotome *Homo erectus* Skeleton. Cambridge: Harvard University Press, pp. 234–265.

SAS Institute (1989) SAS/STAT User's Guide, Version 6, Fourth ed., Volume 2. Cary, NC: SAS Institute Inc.

Sciulli PW, and Giesen MJ (1993) Brief communication: An update on stature estimation in prehistoric Native Americans of Ohio. Am. J. Phys. Anthropol. *92*:395–399.

Sciulli PW, Schneider KN, and Mahaney MC (1990) Stature estimation in prehistoric Native Americans of Ohio. Am. J. Phys. Anthropol. *83*:275–280.

Sjøvold T (1990) Estimation of stature from long bones utilizing the line of organic correlation. Hum. Evol. *5*:431–447.

Sokal RR, and Rohlf FJ (1995) Biometry. Third ed. New York: W.H. Freeman.

StataCorp (1995) Stata Statistical Software: Release 4.0. College Station, TX: Stata Corporation.

St Hoyme L, and Isçan MY (1989) Determination of sex and race: Accuracy and assumptions. In MY Isçan and KAR Kennedy (eds.): Reconstruction of Life from the Skeleton. New York: Alan Liss, pp. 53–94.

Sué E (1775) Sur les proportions du squelette de l'homme. Mem. Acad. R. Sci. *2*:575–585.

Telkka A (1950) On the prediction of human stature from the long bones. Acta Anat. *9*:103–117.

Trotter M, and Gleser G (1952) Estimation of stature from long bones of American Whites and Negroes. Am. J. Phys. Anthropol. *10*:463–514.

Trotter M, and Gleser G (1958) A re-evaluation of estimation of stature based on measurements taken during life and of long bones taken after death. Am. J. Phys. Anthropol. *16*:79–123.

Westfall PH, and Young SS (1993) Resampling-Based Multiple Testing. New York: John Wiley.

Williams RC (1994) Measuring genetic admixture in human populations: The Gila River story. Evol. Anthropol. *3*:84–91.